Bericht TT2P 2

Steffen Brauer, André Harms,
Florian Johannßen, Jan-Christoph Meier,
Florian Ocker, Olaf Potratz,
Torben Woggan

10.06.2012

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Begriffe

Aktion - Belohnung / Reward - Umgebung / Umwelt - Modell der Umgebung - nicht-deterministische Umgebung - stationäre Umgebung - Strategie - stochastische Strategie - deterministische Strategie - optimale Strategie

Return $R_t = r_{t+1} + r_{r_t+2} + ... + r_T$ Return R_t ist Summe der Rewards ab Zeitpunkt t. T wäre ein abschließender Schritt.

Episoden / episodische Aufgaben Handungsabläufe des Agenten können aus Episoden bestehen, die jeweils mit einem abschließenden Schritt enden. Der Agent führt dann episodische Aufgaben durch(episodic tasks).

kontinuierliche Aufgaben Fortdauernde Aufgaben heißen kontinuierlich. Nicht-endende Aufgaben würden zu einer unendlich hohen Belohnung führen. Lösung: Discount-Rate.

terminale / nicht-terminale Zustände

Discount-Rate Nicht-endende Aufgaben würden zu einer unendlich hohen Belohnung führen. Lösung: Abschwächung zukünftiger Belohnungen(discounting). Discount-Rate: $0 <= \gamma <= 1$. Durch Discount-Rate ist Gesamt-Belohnung begrenzt.

- Markov-Eigenschaft, Markov-Zustände, Markov-Entscheidungsprozess - Übergangswahrscheinlichkeit - Erwartungswert für sofortige Belohnung - Zustand-Wert-Funktion (state-value function), V-Wert - optimale V-Funktion - Aktion-Wert-Funktion (action-value function), Q-Wert - optimale Q-Funktion - gierige Strategie (greedy policy) - soft policy, e-soft policy, e-greedy policy - Evaluation vs. Improvement - Strategie-Iteration (policy iteration) - Wert-Iteration (value iteration) - Generalized Policy Iteration - Exploring Starts - On-Policy vs. Off-Policy - Control Problem, Prediction Problem - Behavior Policy, Estimation Policy - Lernrate - Exploration vs. Exploitation - Eligibility - Trace Decay Faktor

2 Lernziele

V1: Wechselspiel zwischen Agent und Umwelt erläutern können V2: Merkmale der Bellman-Gleichungen erläutern können V3: Eigenschaften der Verfahren aus dem Bereich Dynamic Programming erläutern können V4: Verfahren Policy Evaluation, Policy Iteration und Value Iteration erklären können und die Algorithmen in ihren Grundzügen erläutern können V5: Modell der Generalized Policy Iteration erläutern

können V6: Merkmale der Monte Carlo Methoden erläutern können, das Wesen der Verfahren erklären können, Unterschiede zu DP und TD erläutern können V7: Verfahren zur Strategie-Bewertung (First-Visit) und Strategie-Verbesserung (Monte Carlo ES, On-Policy Monte Carlo Control, Off-Policy Monte Carlo Control) in ihren Grundzügen erläutern und vergleichen können V8: TD mit MC und DP vergleichen können V9: Algorithmus TD(0) zur Strategie-Bewertung erläutern können V10: Algorithmen SARSA und Q-Learning zur Strategie-Verbesserung erläutern können, auch die Unterschiede, die sich für Anwendungen ergeben, im Detail beherrschen V11: n-Step TD Predictions und TD(l) Prediction in ihrer Bedeutung erläutern können V12: Eligibility Traces in ihren verschiedenen Ausprägungen mit ihren Vor- und Nachteilen erklären können. V13: Algorithmen TD(l) und Sarsa(l) erläutern können. V14: Grundidee und Problematik des Q(l)-Algorithmus erläutern können

3 Anwendungsaufgaben

A1: Verfahren Policy Evaluation, Policy Iteration und Value Iteration anwenden können, d.h. Berechnungen für gegebene Beispiele durchführen können A2: Das Verfahren zur Strategie-Bewertung „First-Visit“ und das Verfahren zur Strategie- Verbesserung „On-Policy Monte Carlo Control“ anwenden können, d.h. Berechnungen für gegebene Aufgabenstellungen durchführen können A3: Algorithmus TD(0) zur Strategie-Bewertung anwenden können A4: Algorithmen SARSA und Q-Learning zur Strategie-Verbesserung anwenden können